NUANS minihomework 1a

Fabrizio Casadei
ID: 1952529

1 Introduction

In this mini homework, we have faced the problems of **Named Entity Recognition**, unification of Named entities and **Entity liking**.

2 Named Entity Recognition

Utilizing the **Spacy** NLP tool for *NER*, has been defined a pipeline that includes several components like the tokenizer, the POS tagger, the parser, the lemmatizer, the attribute ruler and the fundamental *NER* module. After several trials has been noticed how the pipeline that includes the transformer model (*roberta-base*) trained on web texts, gives more satisfiable and accurate results with respect to the ones of whatever size learned with CNN models. In this mini-homework, we are looking for named entities of three types: Person, Location, and Organization. So a simple filtering between all the entities found has been applied. Moreover has been chosen to consider as Location also the Facilities and the Geo-Political entities, including also these 2 labels under the Location named entities. We end up this step producing a matrix in which each vector contains: the name of the text, start char span, end char span and label associated.

3 Unique Named Entites

From the entities extracted in the NER process, has been defined a simple set of functions to collect under an identical name, the named entities represent the same character, location, or organization. For doing this a data structure has been defined to store together the results from NER which have an equal name. Three different criteria for sorting these information have been implemented: ”longer name”, ”first name”, or ”most frequent name”. Based on the kind of criteria used the unique name associated can change. For the final results has been chosen the ”longer name” criteria. Then another function tries to merge in a ”safe way”, this means merge one entity name that contains or is contained by another named entity, but only if we don’t have an ambiguous case in which the merge with an entity is disputed by two or more other entities.

4 Entity Linking

As last step, we want to associate for each entity a unique identifier of a knowledge base, like the IDs of **BabelNet** (Navigli and Ponzetto, 2012) or **Wikidata**. A code implementation has been provided for both. For retrieving Babelnet IDs, two possibilities have been realized: using **Babelfy** (Moro et al., 2014) or **USeA** (Orlando et al., 2022). Both perform also *Word-sense disambiguation* and are used through the HTTP API. Instead, for *EL* with Wikidata, a new *Spacy* pipeline with CNN models is defined, including the **Spacy Entity Linker** (Martino Mensio, 2017). for both types of *EL*, has been utilized the *edit distance* as a discriminant to associate at each named Entity the corresponding ID. The max distance can be chosen freely (in the results is 2). Before doing this a simple pre-process has been included for the results of NER and the ones of the Entity linking, namely, excluded the English stop words, made lower all the characters, etc. This way of associating the IDs does not find a match for each entity but reduces a bit the possibility of wrong links. it was decided to use the results from Wikidata, since Babelfy free accounts have a limited number of daily queries set to 1000, and USeA takes too much time for the response.

5 Conclusion

This mini-homework gives a simple proof of how it’s easy to handle NLP tools and come out with sufficient results without the need of defining and learn ad-hoc models. Surely for better results, going in-depth with *NER* and *Entity liking* is needed.

References

Emanuel Gerber Martino Mensio. 2017. Spacy entity linker. https://github.com/egerber/spaCy-entity-linker.

Andrea Moro, Francesco Cecconi, and Roberto Navigli. 2014. Multilingual word sense disambiguation and entity linking for everybody. In *ISWC (Posters & Demos)*, pages 25–28.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Riccardo Orlando, Simone Conia, Stefano Faralli, and Roberto Navigli. 2022. Universal semantic annotator: the first unified api for wsd, srl and semantic parsing.